A Prototype System for Retrieval of Gene Functional Information

Lillian C. Folk[1], Timothy B. Patrick, PhD[2], James S. Pattison[3],
Russell D. Wolfinger, PhD[4], Joyce A. Mitchell, PhD[2]
[1]Bioinformatics Consortium, [2]Department of Health Management and Informatics
School of Medicine, University of Missouri-Columbia; [3]Department of Veterinary
Biomedical Sciences, College of Veterinary Medicine, University of Missouri-Columbia;
[4]SAS Institute Inc.

Abstract

Microarrays allow researchers to gather data about the expression patterns of thousands of genes simultaneously. Statistical analysis can reveal which genes show statistically significant results. Making biological sense of those results requires the retrieval of functional information about the genes thus identified, typically a manual gene-by-gene retrieval of information from various on-line databases. For experiments generating thousands of genes of interest, retrieval of functional information can become a significant bottleneck. To address this issue, we are currently developing a prototype system to automate the process of retrieval of functional information from multiple on-line sources.

Background

High-throughput technologies such as microarrays currently allow researchers to gather data about the expression patterns of thousands of genes simultaneously[1]. Statistical analysis techniques and packages are available to rigorously analyze the massive amount of resulting data simultaneously. Yet, though that analysis can reveal significantly expressed genes, in order to make biological sense of those results the researcher is faced with the onerous task of retrieving various types of functional information about them. This task typically degenerates into a manual retrieval of information from various on-line databases, conducted one gene at a time. For large experiments generating hundreds or even thousands of candidate genes of interest, this final step can become a significant bottleneck. To address this issue, we are developing a prototype system to automate information retrieval from multiple on-line sources.

Methods

Interviews with one lab involved in large-scale Affymetrix[2] microarray research were used to generate that particular user's requirements for an automated retrieval system. The resulting system has been implemented, using PERL scripts running under the Linux operating system.

Results

Given a list of Affymetrix probe-set identifiers, the prototype system will: retrieve the target sequences of those probe-sets, submit them to NCBI's online blastn programs, and retrieve the blast results; retrieve the NCBI[3] UniGene, LocusLink, and HomoloGene pages and the Rat Genome Database page (if available); and, from the HomoloGene page, flexibly pursue and retrieve linked pages for homologous genes in the organisms of interest to this researcher (rat, mouse, and human). All processes are performed as batch processes, freeing the researcher for other activity while retrieval is being performed. Currently, all retrieved HTML pages are saved locally. We are currently extending the scripts to extract and summarize particular data (such as Gene Ontology annotations) from the pages.

Discussion

The success of this prototype project supports our belief that intelligent automated retrieval of gene functional information can be made a reality. The current prototype is tailored to the requirements of a single researcher; however, based upon our experience with this system and the results of interviews with other labs, we believe that it could be modified to provide a more generalized, flexible system allowing researchers to perform retrievals according to their own requirements and criteria.

Acknowledgement

We would like to thank Dr. Frank Booth and Dr. Gary Allen, University of Missouri-Columbia, for their support of this project.

References

1. Jacob HJ, Kwitek AE. Rat genetics: attaching physiology and pharmacology to the genome. Nature Reviews Genetics 2002; 3(1):33-42.
2. Affymetrix, [URL: http://www.affymetrix.com/index.affx]
3. Wheeler DL, Church DM, Lash AE, et. al. Database resources of the National Center for Biotechnology Information: 2002 update. Nucleic Acids Res. 2002 Jan 1;30(1):13-6.